Optimizing a physical RNA force-field via Machine Learning

Samuela Pasqual, Laboratoire CiTCoM
Frédéric Lechenault, LPENS

Motivations

In humans only 1-2% of DNA codes for proteins, most of the rest codes for RNAs. Most of these RNA molecules are involved in gene regulation mechanisms. Unveiling their structures and functions will allow to expand drug design to RNA targets, opening to a whole new branch of medicine. Most viruses heavily rely on RNA, either having an RNA genome or having RNA molecules perform enzymatic and regulatory activities. Understanding these mechanisms will allow to interfere with their life cycles and stop infections.

Problem statement

We developed a coarse-grained physical model for RNA molecules [1] that allows in principle to predict folding and follow conformational changes. This model needs to be optimized finding the correct sets of parameters to describe the interactions.

The model functional forms have been chosen empirically. With the aid of ML we will also be able to learn possible corrections to the functional forms used, therefore learning the force-field itself from the data.

Proposed solution

Use ML to find the optimal set of parameters of the force-field (100+) to distinguish native structures from decoys, i.e. train the system to distinguish correctly folded structures.

- Set up the global optimization scheme coupling Pythorch and the coarse-grained model.
- Generate the appropriate set of training structure and decoys.
- Run the optimization on the training set.
- Run simulations on benchmark systems using the new parametrization.

Related Work

The project stems from the synergy of two groups, one working on RNA modeling the other on Machine Learning applied to physical systems.

Concerning RNA modeling, we work both on developing new simulations methods [3,4] as well as to applications to specific systems in collaboration with experimental teams [5].

Concerning Machine Learning, we work at developing an interface between physics experiments and ML techniques aimed at gathering insights into the mechanisms at hand in situations such as earthquakes predictions, active non-linear control or DNA decoding [6,7].

RNA modeling

Symbolic Regression algorithms

Dataset Description

The main source of data available are the high-resolution experimental structures deposited in the Nucleic Acids Data Bank, containing several thousands molecules.

Through simulations, for these systems it is possible to generate unfolded, partially unfolded, or folded but competing structures to be used as decoys [2].

Other sources of data come from thermodynamics predictions or data reported with the deposited structures, such as ionic concentrations and pH.

Results

We expect to have two sets of successive results:

1. Short-term: A newly parametrized force-field to based on the current model to be used in the very near future to simulate RNAs of high pharmacological significance such as G-quaduplexes (involved in cancer regulation) and the frameshifting element of SARS-CoV-2.
2. Long-term: A new force-field where the physical interactions will be deduced by ML-based symbolic regression, to be integrated in a highly parallelized code in development in our groups and that will be made available to the whole scientific community.

References

1. T. Cragnolini, Y. Laurin, P. Derreumaux, S. Pasquali, **The coarse-grained HiRE-RNA model for de novo calculations of RNA free energy surfaces, folding, pathways and complex structure predictions**, J. Chem. Theory Comput., 11, 3510 (2015)
2. K Röder, S Pasquali, **RNA modeling with the Computational Energy Landscape Framework**, RNA Scaffolds, 49-66 (2021)
3. L Mazzanti, L Alferkh, E Frezza, S Pasquali, **Biasing RNA coarse-grained folding simulations with Small--Angle X--ray Scattering (SAXS) data**, Journal of Chemical Theory and Computation, accepted for publication (2021),
4. S. Pasquali, E. Frezza, F.L. Barroso da Silva, **Coarse-grained dynamic RNA titration simulations**, Interface Focus 9: 20180066 (2019)
5. K Röder, G Stirnemann, AC Dock-Bregeon, DJ Wales, S Pasquali, **Structural transitions in the RNA 7SK 5' hairpin and their effect on HEXIM binding,** Nucleic Acids Research 48 (1), 373-389 (2020
6. Lechenault F, Baker A, Krzakala F**, Deep seq2seq architecture for DNA sequence decoding from noisy data** NeurIPS proceedings (2019)
7. Adèle Douin, Frederic Lechenault, Jean-Philippe Bruneton, **Machine Learning Predictions of Avalanche-like Events in Knitted Fabric**, Bulletin of the American Physical Society (2021)

This work is supported by diiP, IdEx Université de Paris, ANR-18-IDEX-0001